Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W295–W298***
doi:10.1093/nar/gki406

Manipulating multiple sequence alignments via MaM and WebMaM

Can Alkan*, Eray Tüzün[1], Jerome Buard[2], Franck Lethiec[3], Evan E. Eichler[1], Jeffrey A. Bailey[4] and S. Cenk Sahinalp[5]

Department of EECS, Case Western Reserve University, Cleveland, OH, USA, [1]Department of Genome Sciences, University of Washington, Seattle, WA, USA, [2]Institute of Human Genetics, CNRS UPR 1142, Montpellier, France, [3]Mèthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS UMR 2506, Montpellier, France, [4]Department of Genetics, Case Western Reserve University, Cleveland, OH, USA and [5]Department of Computing Science, SFU, Burnaby, BC, Canada

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

MaM is a software tool that processes and manipulates multiple alignments of genomic sequence. MaM computes the exact location of common repeat elements, exons and unique regions within aligned genomics sequences using a variety of user identified programs, databases and/or tables. The program can extract subalignments, corresponding to these various regions of DNA to be analyzed independently or in conjunction with other elements of genomic DNA. Graphical displays further allow an assessment of sequence variation throughout these different regions of the aligned sequence, providing separate displays for their repeat, non-repeat and coding portions of genomic DNA. The program should facilitate the phylogenetic analysis and processing of different portions of genomic sequence as part of large-scale sequencing efforts. MaM source code is freely available for non-commercial use at http://compbio.cs.sfu.ca/MAM.htm; and the web interface WebMaM is hosted at http://atgc.lirmm.fr/mam.

INTRODUCTION

The construction and analysis of multiple sequence alignments of genomic DNA is central to phylogenetic and comparative sequencing initiatives. Although there are many programs that can compute multiple alignments of genomic sequences [e.g. Clustal W (1), MSA (2), HMMER (3), etc.], and methods, such as MLAGAN (4) and MAVID (5) first find locally conserved regions and then combine these anchors into larger alignments, none are capable of discriminating various biological elements within the given alignment. The phylogenetic analysis of aligned genomic DNA is complicated by the fact that DNA is heterogeneous in nature. It is composed of differing fractions of protein-encoding regions, common repetitive elements, regulatory regions and satellite DNA. These different regions have been shown to be subject to differing processes of mutation and selection. For example, effective rates of sequence variation may differ dramatically between exons and flanking intron sequences owing to forces of purifying and to a lesser extent positive Darwinian selection operating on non-synonymous nucleotide base pair positions within coding exons. Common repeat elements are frequent targets for gene conversion events which obscure the evolutionary relationships between orthologous segments of the genome. Similarly, tandem repeats, such as microsatellites, minisatellites and satellite DNA have been shown to mutate by non-stepwise models of mutation, resulting in extraordinary rates of evolutionary turnover, making comparative alignment, even among closely related species of such regions, problematic. Failure to resolve such heterogeneity may lead to erroneous estimates of divergence between genomic regions or create phylogenetic relationships among sequences which can not be directly compared with earlier studies, which have focused largely on coding and/or non-repetitive regions of the genome.

Recapitulation of the phylogenetic relationships among genomic sequence requires that this heterogeneity be recognized and the segments be treated independently. Traditional methods for parsing such regions from large-scale genomic sequence can be a tedious and time-consuming task. With this motivation we developed MaM, which is a powerful tool for analyzing and parsing multiple sequence alignment files. The currently available version (MaM 1.2, available for download) is suitable for parsing alignment data, finding the common repeat locations, exonic regions, etc., extracting these regions to generate subalignments for independent analysis and for

*To whom correspondence should be addressed. Tel: +1 604 2687040; Fax: +1 604 2913045; Email: calkan@cwru.edu displaying variations within these alignments. MaM supports common phylogenetic formats as standard input/output, such as Clustal (1), NEXUS (6), MEGA (7) and FASTA (8).

Further, the tabular format of MaM provides a flexible platform for the treatment of other sequence motifs and properties that may be recognized by other programs and require to be analyzed independently.

METHODS

Input and output files

MaM requires a single text file containing the alignment of multiple sequences. The input formats supported by MaM are Clustal (1), NEXUS (6), MEGA (7) and Fasta (8). After the execution, MaM creates three different output files: (i) a postscript file that contains a graph displaying both the nucleotidic divergence between sequences and the location of DNA elements, such as repeats, unique sequences or exons along the alignment (see 'Displaying the alignment property' section and Figure 1 for more detail), (ii) a text file that contains the merged sequence of either the repeats or the unique sequences found within the alignment and (iii) a text table file that contains the begin and end locations of each sequence element that is included in the merged sequence. The input/output formats and user switches used in MaM will be explained in detail further in the paper.

Finding sequence motifs

Given a multiple sequence alignment file, which involves some n sequences, each of size m, MaM first extracts the gapped version of each sequence from the alignment file. After this step, MaM removes all gaps from each gapped sequence to obtain its original, non-gapped version. Then, MaM gives the option of applying one of the repeatmasker (http://www.repeatmasker.org), cross_match (http://phrap.org) or sim4 (9) programs to each of the sequences. It is also possible to import an exonfile, which should be a cDNA sequence file, to use with one of these programs. In that case the motifs extracted by the selected program would be the locations of exons instead of repeats. This step provides partial information about the locations of sequence motifs (provided by the program chosen by the user) in each non-gapped sequence in the dataset. This information is kept as a table of these sequence motif locations (begin and end positions on the non-gapped version) for each sequence. The user is also given the choice of selecting tablefile option instead of the three other programs. If done so, the user imports own table of motif locations, thus bypassing the previous step. The location information is then converted to that for the gapped (i.e. aligned) version for each sequence.

After this step MaM will have n tables for each gapped sequence with begin and end locations of each sequence element. MaM then merges these n tables into a single table that summarizes the repeat locations for these n sequences. Here MaM provides the option of performing the merge by either intersecting the sequence elements or concatenating them. We describe how this is performed in more detail below.

Basic steps for finding sequence motifs

Input. Alignment file with n sequences, $S_1, S_2, \ldots, S_n$.

Output

(i) $T_{o,i}$: Table file for each non-gapped sequence S_i indicating the begin and end locations of all repeats.
(ii) $T_{a,i}$: Table file for each gapped sequence $S_{i,}.aln$ indicating the begin and end locations of all repeats.

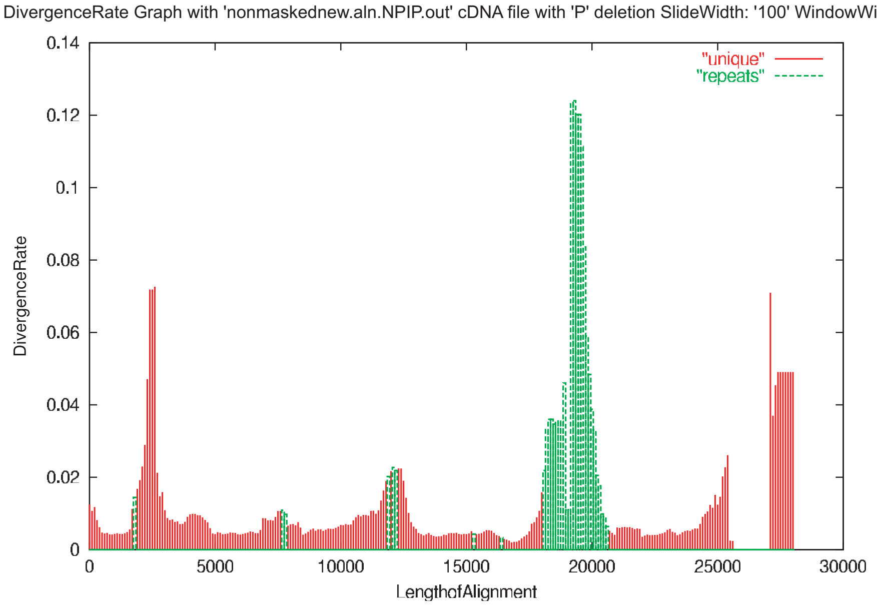

Figure 1. MaM graphical display of divergence rate across a multiple alignment. Seventeen genomic segments each corresponding to a ∼20 kb region were aligned using Clustal W. Variation in sequence divergence was measured over 1 kb windows sliding every 100 bp. The positions of exons (marked as 'repeats', shown in green; whereas introns are marked as 'unique' and shown in red) were mapped using MaM and correspond to the morpheus gene family. Some of the exons show evidence of accelerated rates of nucleotide variation when compared with the intronic regions.

Algorithm

```
Parse the alignment file to obtain the gapped sequences $S_1.aln$, $S_2.aln$, ..., $S_n.aln$.
For every gapped sequence $S_i.aln$, obtain the non-gapped sequence $S_i$ by removing all spaces.
If table file is not given by the end user, then for every sequence $S_i$ execute one of cross_match/repeatmasker/sim4 to obtain a $T_{o,i}$.
Otherwise, process table file to obtain $T_{o,i}$.
Convert every table file $T_{o,i}$ to $T_{a,i}$.
Start by assigning $T_{a,final}$ to $T_{a,1}$; for every sequence $S_i$ $(i > 1)$,
    merge $T_{a,final}$ with $T_{a,i}$ according to either MAX (i.e. concatenation) or MIN (i.e. intersection) criteria.
```

Procedure merge. Given table files $T_{a,final}$ and $T_{a,i}$, the merging of the sequence elements in the tables can be done by first sorting the sequence elements from both tables into a single table. This table, however, will have overlaps between the motifs. These overlaps can be eliminated by going through the table from left to right, according to either MAX or MIN criteria chosen by the user.

Under MAX criterion MaM computes the union of two successive overlapping motifs by setting (i) the smallest of the begin locations as the begin location of the concatenation and (ii) the largest of the end locations as the end location of the concatenation.

Under MIN criterion, MaM computes the intersection of two successive overlapping motifs by setting (i) the largest of the begin locations as the begin location of the concatenation and (ii) the smallest of the end locations as the end location of the concatenation. We demonstrate how this works on an example in Table 1.

Displaying the alignment

After computing the begin and end locations of all sequence elements, the user can display the alignment property by a sliding window method. The user also has the option to select all sequence motifs (determined in the previous step) to be displayed. This is done as follows.

Cutting and displaying sequence motifs in the alignment. If the user decides to cut the elements from the alignment MaM provides two further options: (i) displaying each motif separately or (ii) concatenating all motifs and displaying them together.

The first option is self explanatory. For the second option, consider, for example, an alignment in which there are two sequence motifs, as given in Table 2. The output of this option will be concatenation of the two motifs displayed as a single sequence, the character #250 is immediately appended to character #200 in the motifs.

Table 1. Merging sequence elements

$T_{a,final}$		$T_{a,i}$		T_{merged} MAX		T_{merged} MIN	
30	50	20	40	20	50	30	40
60	80	70	80	60	80	70	80

Table 2. Cutting and displaying sequence motifs in the alignment

Motif 1		Motif 2	
Begin: 1	End: 200	Begin: 250	End: 400

Displaying the alignment property. MaM displays the quality of alignment by sliding a window (whose size is determined by the user) through the whole alignment. MaM computes the divergence score for the alignment confined in each window position and displays how this score varies over the whole alignment. For calculating the divergence score within a window, the user has three choices. The first option is using sum-of-pairs score (denoted as pairwise deletion in the program), where every pair of bases in the same column of the alignment are compared with each other and the ratio of diverging pairs (not counting the gaps) are reported. The second option is the same except all the columns that contain a gap/deletion from the alignment are ignored (denoted as complete deletion). The last option is denoted as parsimony score and the ratio of the bases that differ from the mostly occuring base in the same column is returned. Examples for all three options are given in the MaM website. The display of alignment quality also highlights the sequence elements determined in the previous step. An example graph generated by MaM by using gnuplot (http://www.gnuplot.info) with 17 sequences of size 20K is given in Figure 1.

Web Interface for MaM

We also developed a web interface for MaM at Montpellier Laboratory of Computer Science, Robotics and Microelectronics (WebMaM: http://atgc.lirmm.fr/mam). The web version currently implements limited functionality of MaM (see Methods section for full description). WebMaM runs only repeatmasker to find sequence motifs (other options are not available in the current version); exonfile upload is not allowed, and complete deletion switch is preset. We are still in the development process for this web version, and full functionality will be available in the near future.

The input file for WebMaM is the same as with MaM, the other options can be set through the web interface. After the process, three output files are sent to the user via e-mail.

DISCUSSION

Implementation and performance

MaM is written in C and runs on both Linux and Unix platforms. Given n aligned sequences each of size m, the memory requirement of MaM is $O(n{\cdot}m)$. We observed that on a 2 Ghz Pentium IV PC with 1 GB RAM, MaM can successfully perform the desired computations for input sequences of size n and m, where $n \times m = 1$ billion. Although the running time of MaM depends on both n and m, its main bottleneck is the execution of the programs for computing the repeat elements.

MaM requires the installation of repeatmasker (http://www.repeatmasker.org), cross_match (http://www.phrap.org), sim4 (9) packages and gnuplot program (http://www.gnuplot.info).

ACKNOWLEDGEMENTS

MaM is developed at Center for Computational Genomics, CWRU; supported in part by grants from Charles Wang Foundation and the Ohio Board of Regents. Web interface for

MaM is developed at Montpellier Laboratory of Computer Science, Robotics, and Microelectronics (LIRMM) thanks to the support of the Genopole. J.B. is supported by the CNRS ACI IMPBio. Funding to pay the Open Access publication charges for this article was provided by NSERC, Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
2. Lipman,D.J., Altschul,S.F. and Kececioglu,J.D. (1989) A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA*, **88**, 4412–4415.
3. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
4. Brudno,M., Do,C., Cooper,G.M., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*., **13**, 721–731.
5. Bray,N. and Pachter,L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res*., **14**, 693–699.
6. Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics, 2nd edn*. Sinauer Associates, Sunderland, MA, pp. 407–514.
7. Kumar,S., Tamura,K., Jakobsen,I.B. and Nei,M. (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. *Bioinformatics*, **17**, 1244–1245.
8. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
9. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*., **8**, 967–974.